Statement of research interest

Milos Hauskrecht

My primary field of research interest is Artificial Intelligence (AI). Within AI, I am interested in problems related to probabilistic modeling, machine learning and data mining, and their interdisciplinary applications to biomedical sciences. I have found the applications in biomedical sciences to be exciting not only because of their practical value, but also as an inspiration for new types of problems that prompt the development of new theories and algorithms that have not been investigated before. As a result, my research work blends both theoretical and practical aspects of problems.

In the following, I first briefly summarize research work I conducted till approximately year 2009. This work also includes my early career years as a PhD student at MIT, and a postdoc at Brown University. After that I describe in more depth research areas I currently investigate with my students and collaborators.

A. Research Work till 2009

My research work during this period focused mostly on two areas: (1) planning and decision-making in the presence of uncertainty, and (2) machine learning and data mining. Both of these areas were strongly influenced by biomedical and engineering problems.

Planning and decision-making in the presence of uncertainty. This line of my research work focused on the development of Markov decision process (MDP) models and algorithms for their optimization. Briefly, Markov decision processes provide an elegant mathematical framework for modeling the dynamics of the environment in response to various actions/interventions. The models are useful for decision analysis and planning, where the goal is to identify the best sequence of actions to take in order to optimize some outcome measures (such as, quality-adjusted life-years in medicine). Many different variants of the basic Markov decision process model exist. Typically, the more expressive and flexible these variants are, the harder they are to solve, and clever approximations are needed to solve them efficiently. My key contributions to this field are in the development of such approximation methods and their applications. Examples of my work include: (1) my dissertation research on efficient upper and lower bound approximation methods for Partially observable Markov decision processes (POMDPs) and their application to the problem of modeling and optimizing the management of patients with ischemic heart disease; (2) the development of hierarchical decomposition methods for solving large MDP problems more efficiently, and (3) the development of Approximate Linear Programming (ALP) methods for solving large factored MDPs with continuous or hybrid state and action spaces.

Machine learning and data mining. Motivated by the emergence of large datasets in various areas of science, technology, business, and everyday life in the past two decades, my second area of research work focused on machine learning and data mining that aim to improve the understanding of patterns in the data and processes that generate them, and utilize them in solving nontrivial and challenging problems such as diagnosis, prediction of outcomes, or event detection. My work spanned both the development of new machine learning algorithms, as well as, their interdisciplinary applications to biomedical and engineering fields. Examples of my work in biomedical sciences include: (1) the development of machine learning methods for analysis of high-throughput genomic and proteomic profiles, construction of multivariate models for early-detection and diagnosis of disease, and their application to analysis of case/control proteomic datasets related to: pancreatic cancer (UPCI), lung cancer (UPCI, Vanderbilt) melanoma cancer (UPCI), lung disease (UPMC), kidney transplant (UPMC), dental caries (NIH) diabetes (Harvard), childhood arthritis (UPMC) and scleroderma (UPMC); and (2) the development of a novel latent variable framework to identify key regulatory signals in high-dimensional gene expression data, which we successfully tested on yeast cell-cycle data from the Stanford gene expression database. Examples of my other machine learning work include (1) the development of the Noisy-or component analysis model for representing interactions among entries in high-dimensional binary datasets, which we applied to analysis of co-citation data in Citeseer documents and to modeling of traffic congestion patterns; and (2) the modeling of high-dimensional continuous distributions based on the Mixture of Gaussian trees model and approximate algorithms for their efficient learning and inference, which we applied to modeling of vehicular traffic flows.

B. Current research work

My current research work focuses on the design of new machine learning and data mining methods for analysis and utilization of complex large-scale time-series datasets derived from Electronic Health Records (EHRs). This direction builds upon and naturally combines experience from my previous research work in developing Markov models of dynamics, and machine learning methods for complex high dimensional data. It also builds upon my past experience in biomedical informatics and biomedical applications.

In the following, I describe five main directions I currently investigate with my students and my collaborators. These efforts have been funded by four NIH grants on which I am a PI, one NSF grant and one NIH grant on which I serve as a co-investigator. The grants are:

- **NIH.** 2R01GM088224. Real-time detection of deviations in clinical care in ICU data streams. (PIs: Hauskrecht and Clermont), August 2014- June 2018.
- **NIH.** R01LM011966-01. Improving Clinical Decision Support Reliability Using Anomaly Detection Methods. (PI: Adam Wright, Partners healthcare, Boston, MA), July 2014- June 2018.
- **NSF.** IIS 0911032. Discovering Complex Anomalous Patterns (PI: Dubrawski, CMU), September 2009-August 2014.
- **NIH.** 1R01GM088224. Detecting deviations in clinical care in ICU data streams. (PIs: Hauskrecht and Clermont), September 2009 - June 2013.
- **NIH.** R21LM009102. Evidence-based anomaly detection in clinical databases. (PI: Hauskrecht), April 2007-April 2009.
- **NIH.** 1R01LM010019. Using medical records repositories to improve the alert system design. (PI: Hauskrecht), September 2009 - September 2013.

B.1. Outlier-based monitoring and alerting

Detection and prevention of unusual events is an important issue in highly interconnected and computerized environments, mostly because it is demanding both in terms of human labor and capital investment. Statistical anomaly detection provides a set of techniques that are capable of identifying rare (or in other words, anomalous) events in the context of large environments (population of patients, entries in a database, etc.). Our objective is to utilize anomaly detection in clinical monitoring and medical error detection.

The goal of this project funded by NIH grants 2R01GM088224 (PIs: Hauskrecht and Clermont), 1R01GM088224 (PIs: Hauskrecht and Clermont) and 1R21LM009102 01A1 (PI: Hauskrecht), is to develop and evaluate data-driven, rather than expert-based, solutions to detect anomalous patient-management actions and use them to generate clinical alerts. Our approach works by identifying patient-management actions that are unusual with respect to actions used to manage comparable patients in the past and by raising a patient-specific alert when such an action is encountered prospectively. The main hypothesis we investigate is that statistical anomalies in patient management actions correspond to medical errors often enough to justify the alerting.

Classical approaches in detection of medical errors utilize a set of expert-defined rules to detect anomalous events. Apart from the fact that these systems often lack well-defined statistical underpinning, adaptation of the systems to a new environment is impossible without direct interaction with a human. On the other hand, statistical methods rely on past data as opposed to carefully extracted expert knowledge. The benefit of the statistical approach is that it is adaptive and can be applied to a wide range of conditions. In summary, the new outlier-based alerting approach is designed to complement existing knowledge-based detection and alerting methods that are clinically precise, but costly to build.

Conditional anomaly detection. The majority of existing anomaly detection methodologies focus on detection of anomalous data entries in the datasets. This biases anomalies to events that occur with a low prior probability, for example, patients suffering from a rare disease or exhibiting a rare combination of symptoms. In our work, we are interested in finding anomalies in outcomes or patient management decisions with respect to patients who suffer from the same or similar condition. This problem is referred to as conditional (or contextual) outlier detection. To account for the conditional aspect of anomaly detection we have developed and continue refining conditional anomaly detection methods that seek to detect unusual values for one or a subset of variables (outcomes, decision) given the values of the remaining variables. We have developed and studied a number of conditional outlier detection methods based on discriminative probabilistic classifiers, nonparametric graph-based methods, and instance-based classifiers. Another challenge for building anomaly detection methods for EHRs is that data are complex multivariate data. This opens up an important question of how to represent the patient and patient's state efficiently. I discuss the approaches to do so later in Section B.3.

Evaluation of the framework. We have developed and refined a new outlier-based alerting methodology that works with clinical time-series and have tested it on structured entries in EHRs stored in past patient data archives. We have conducted two evaluation studies: first one used EHRs of 4468 post-surgical cardiac patients, and the second one used EHRs for over 25000 ICU admissions. The evaluation studies used critical care physicians (15 and 18 physicians for the two studies respectively) who evaluated the alerts raised by our system, and the results showed that the methods reach true alert rates above 50%, that is, one correct alert in two alerts raised by the system. These numbers are very promising when compared to true alert rates of clinical and drug alerting systems reported in the literature that range from 2% to 52%. Our work on outlier-based alerting was recognized by the Homer R. Warner Award during the American Medical Informatics Association Annual Symposium in Washington DC in November of 2010.

Current and future work. Our current research on outlier-based alerting concerns integration of the system with the real-time hospital production system and the refinements of the approach to multivariate outliers. We believe the outlier-based monitoring and alerting approach we are pursuing in our research has great potential for detecting medical errors that would otherwise go unnoticed. We have high hopes this approach and the systems based on it will be eventually deployed at the bedside and we envision its future integration with existing knowledge-based alerting systems.

Finally, in collaboration with Adam Wright from Brigham and Woman Hospital in Boston, we are starting a new (recently funded) NIH project (R01LM011966-01) that aims to design and study computational methods for screening the operation of Clinical Decision Support (CDS) systems that will be able to identify, as early as possible, their malfunctions.

B.2. Predictive pattern mining and subgroup discovery

This area of our work, funded in part by NIH grants 1R01LM010019-01A1 (PI: Hauskrecht), 1R01GM088224 (PI: Hauskrecht and Clermont) and NSF IIS0911032 (PI: Dubrawski), has focused on subgroup discovery, which is the problem of identifying patterns in data that are most important for predicting and explaining a specific outcome variable. An example is the problem of finding subpopulations of patients that respond better to a certain treatment than the rest of the patients. With the emergence of large datasets in all areas of science, technology and everyday life, identification of predictive patterns characterizing different subgroups is extremely promising for: (1) knowledge discovery purposes, especially when new, previously unknown, subgroups (e.g. subpopulations of patients) with significantly different outcomes are identified; and (2) feature engineering, when patterns found are used as features one may include when building various classification models to predict the outcome variable.

Minimal predictive pattern mining. Standard predictive pattern mining algorithms work by scanning many different patterns for their ability to predict the outcome variable and assess their quality in terms of some predictive score. The predictive patterns these algorithms return include the patterns that satisfy the pre-specified predictive score threshold. However, many predictive patterns selected this way are redundant in that they do not bring any (or very little) new information when compared to more general patterns that represent larger populations and that were already included in the result. To address the above problem we have developed the minimal predictive pattern mining framework that eliminates pattern redundancies by selecting only those discriminative patterns that are significantly different from more general patterns. To measure the redundancy of a predictive pattern, we have developed and tested two criteria based on the binomial score and the Bayesian score that assess the chance of that pattern being consistent with a more general pattern. We showed that the minimal predictive pattern mining algorithm leads to a more compact description of the predictive differences among subgroups which is beneficial for knowledge discovery. We have also shown that minimal predictive patterns improve the accuracy of classification models based upon these patterns.

Temporal predictive pattern mining. One of the key factors motivating this area of our work was the application of the methods to analysis and explanation of predictive patterns in clinical data that are recorded in EHRs. This is extremely challenging since EHRs consists of complex multivariate time series of observation, tests, and treatments. To construct temporal patterns from complex clinical data we rely on the temporal abstraction approach to obtain a high-level qualitative description of the time series. The temporal abstractions are then combined with temporal logic to form more complex temporal patterns for the abstracted data. This representation allows us to express temporal concepts such as ”the administration of heparin precedes a decreasing trend in platelet counts” representing the different patient subpopulations and it can be integrated within the minimal predictive pattern mining framework described above.

Recent temporal predictive patterns. The space of possible temporal patterns one can define on time series data with the help of temporal abstractions is enormous. While the minimal predictive pattern mining framework helps us to reduce the number of patterns the algorithm finds, it may still search and scan through a very large number of patterns. The key challenge is to find ways of reducing the complexity of this space as much as possible, hence improving the efficiency of the mining algorithms. To address this concern we proposed a new approach that builds predictive patterns for monitoring and event detection problem using the ’recent predictive pattern’ heuristic which captures the intuition that most recent information related to the clinical variable is likely to be the most important for future prediction. We successfully tested this approach by predicting: (1) patients who are at risk of developing heparin induced thrombocytopenia (HIT), and (2) complications (diagnosis of secondary diseases) for diabetic patients.

B.3. Modeling of clinical time series and construction of temporal features

The key challenge for analysis of clinical data is that EHRs consist of complex multivariate time series of clinical variables collected for a specific patient, such as laboratory test results, medication orders, physiological parameters, past patient’s diagnoses, surgical interventions and their outcomes. In addition, this heterogeneous time series may also vary in length depending on the age of the patient or the length of patient’s hospitalization. A fundamental challenge the researchers and data analysts face is how to summarize and represent this complex time-series data in order to make them amenable to statistical analysis and modeling. Our work has focused primarily on predictive modeling problems and solutions where our goal is to identify or construct temporal features important for prediction of adverse events, patient outcomes, and future patient management decisions. We have studied three different approaches to this task which I briefly review next.

Fixed feature maps, The first approach we have developed to address the above problem relied on a fixed set of feature mappings defined by domain experts that allows the conversion of complex time series of labs, medications or physiological values to different temporal features. For example, the time series of values for a numerical lab (such as platelet count) were converted to 28 different features reflecting the last lab value observed, the time elapsed since the last value, last trend, apex and horizon values, and their differences from last value, etc. We have developed similar feature sets for medications orders (e.g. aspirin or heparin), physiological parameters, input/output volumes, etc. We have successfully used these feature sets in modeling the patient state for adverse event detection, and in the outlier based alerting projects.

Temporal pattern features. While expert-defined feature set mappings turned out to be extremely useful for multiple projects, one wonders if there are other temporal features important for prediction tasks one should include and whether these can be constructed automatically from data. One approach we pursued to address this problem was based on the ideas of predictive pattern mining and predictive temporal pattern mining discussed in the previous section (Section B.2). Briefly, predictive pattern mining approach addresses the feature construction problem by extracting predictive patterns characterizing patient subgroups in EHRs that are important for prediction. Such patterns, when explicitly represented, can be used as features one needs to include when defining classification models. As described in the previous section our work focuses on temporal patterns based on temporal abstraction and temporal logic to get a high-level qualitative description of clinical time series. This representation is more flexible and allows us to express various temporal patterns observed in clinical time-series data.

State-space models of dynamics. Another direction we are currently exploring attempts to build models of multivariate time series data and their behaviors by exploring lower-dimensional representations of the patient state with the help of various Markov process models. Briefly, our goal is to find a lower dimensional patient state representation that summarizes and compactly encodes all information about past patient's observations that is needed to predict well the behavior of the time series in the future. Such a state representation can yield a compact feature space for various prediction problems in clinical time series and would be an alternative to features based on fixed feature mappings or predictive patterns discussed earlier. We have studied and developed models based on Linear dynamical system and Gaussian processes frameworks, and their hierarchical combination to model time-series of the different laboratory tests. We currently continue expanding these models to handle observations from multiple labs defined by multivariate time-series.

B.4. Cost-effective labeling of data by human experts

Learning of classification models in medicine often relies on data that are labeled or annotated by a human expert. Since labeling of clinical data may be time-consuming and costly finding ways of alleviating the labeling costs is critical for our ability to automatically learn such models from data. The development of new methods that reduce the dependency on the number of labeled examples becomes critical for practical application and deployment of such models. Our recent work has focused on and investigates several approaches to address the labeling bottleneck.

Active learning/sampling. One of the most popular research directions for reducing the labeling cost is active learning. The goal of active learning research is to develop methods that analyze unlabeled examples, prioritize them and select those that are most critical for the task to be solved, while optimizing the overall data labeling cost. In our work, we have studied a special active learning (sampling) framework, that leads to a labeled set of patient cases that can help us to evaluate, as accurately as possible, various statistics of alerting rules and predictive models. This framework is the basis of our NIH funded work (grant 1R01LM010019-01A1 PI: Hauskrecht), that aims to aid clinicians in the construction of new rules and in the refinement of existing alerting rules by providing early (offline) feedback on their expected performance based on EHR data stored in patient archives. The ability to conduct early offline evaluation of alerting rules helps to shorten otherwise time-consuming rule design and rule refinement cycle.

Auxiliary soft label information. Another direction we are pursuing to alleviate the labeling bottleneck is the use auxiliary information to construct a predictive model. We have proposed and developed a new machine learning framework in which the binary class label information that is typically used to learn binary classification models is enriched with soft-label information reflecting more refined experts view on the class a labeled instance belongs to. In general, the soft label information can be represented either in terms of (1) a probabilistic (or numeric) score, e.g., the chance of the patient having the disease is 0.7, or, (2) a qualitative category, such as, weak or strong agreement with the patient having the disease. The cost of obtaining this additional information is typically small once the patient case is reviewed by the expert. We have proposed a number of new algorithms and models that can accept this information in the model training phase, and demonstrated that they can help us to learn classification models more efficiently (with a smaller number of labeled examples) than with binary labels only.

Learning from multiple experts. Standard machine learning framework assumes that the labels are assigned by a homogeneous process. However, in reality the labels may come from multiple experts and it is possible that the experts may have different subjective opinions on how some of the patient cases should be labeled. We studied this scenario by designing a new multi-expert learning framework that assumes the information on who labeled the case is available. Our framework explicitly models different sources of disagreements and lets us naturally combine labels from different human experts to obtain: (1) a consensus classification model representing the model the group of experts converge to, as well as, (2) individual expert models. We have tested the proposed framework by building a model for the problem of detection of the heparin induced thrombocytopenia (HIT) where the cases are labeled by three experts. We showed that our framework is superior to multiple baselines (including standard machine learning framework in which expert differences are ignored) and that our framework leads to both improved consensus and individual expert models.

Current work. A number of other approaches for alleviating the labeling effort are possible. We are currently studying special transfer and semi-supervised learning approaches one can apply to time-series data with the aim of reducing the labeling effort for event detection tasks. Our expectation is that by explicitly modeling the new information sources and by efficiently learning their relations to event labels we will be able to accurately infer (fill in) the missing labels for many classification tasks.

B.5. Non-parametric graph-based methods

The notion of similarity among data objects plays a fundamental role in many machine learning methods. Graph-based methods induce similarity between data objects by first forming a similarity graph from local similarities, and then performing spectral decomposition on the graph which aims to aggregate the effects of local similarities into a global (data-driven) similarity metric (or kernel) between the objects. A direct consequence of this is that machine learning algorithms that are designed to work with absolute distances can be now applied to problems with data-driven distances. Our research on the graph-based methodology has covered: (1) the development of approximation methods for large-scale, high-dimensional data; (2) the development of a new text metric to support inferences in text; (3) application of the methodology to support inferences in clinical data.

Approximation algorithms for large-scale and high-dimensional data. One of the challenges of graph-based methods is their scalability to high-dimensional and large-scale data. We worked on two approaches to address these challenges. First, we developed a variational dual-tree framework to approximate the transition matrix of the random walk on the graph (or equivalently the random walk Laplacian matrix). This transition matrix is specifically useful in applications such as diffusion analysis, semi-supervised learning and link analysis. We demonstrated an order of magnitude speedup for label propagation tasks on real-world datasets with this framework. Second, we proposed a novel approach for building the kernel for high-dimensional data by factoring it into to multiple kernels defined on indepenendent data subspaces.

Graph-based text metrics. Graph-based methods can be used to induce similarities for an arbitrary set of objects from their local similarities. We pursued this approach by defining similarities among terms in a document corpus. Briefly, we used pairwise co-occurence counts of terms in sentences and paragraphs of documents to define their local similarity. The global similarity was then induced by using spectral decomposition methods to derive smooth (local-similarity preserving) kernels. This approach lets us build similarities among pairs of terms or concepts. After that we developed a novel way of defining similarity among sets of terms which let us compare sentences, paragraphs or even whole documents. To compute this generalized text similarity, we developed an approximation technique that relies only on the set of precompiled pairwise term-term similarities. We demonstrated the benefits of the entire framework on three text inference tasks: (1) prediction of terms in a biomedical article from its abstract and (2) query expansion in information retrieval, and (3)a gene prioritization task, where our goal was to infer from the text the genes that are most likely associated with Alzheimer disease.

Applications to clinical data. Our current research in non-parametric graph-based methods focuses primarily on the application of the methodology to clinical data, where our goal is to define similarity among patients and their clinical time series. Our initial work in this direction used graph-based methods in combination with fixed feature maps outlined in Section B.3. to detect outliers in patient management actions.